MAY 2026

AMERICANS for RESPONSIBLE INNOVATION

Hands on the Trigger

The Operational Necessity of Humans with Lethal Autonomy

WHITEPAPER BY:

JACQUELINE VITZENIK

Executive Summary

Artificial Intelligence (AI) and its use of warfare have dominated public discussions ever since OpenAI's ChatGPT launched in 2023, and even more with recent reporting of Anthropic's Claude use in Operation Absolute Resolve[1] and the Iran war[2] via the Maven Smart System. Much of this debate has focused on the implications of fully autonomous weapon systems capable of using lethal force without any human supervision, let alone intervention. However, these concerns and these weapon systems are not new. The most urgent current issue is not autonomy per se, but the integration of modern AI/ML into a wider set of operational contexts than earlier, more bounded autonomous systems. This is a categorically new phase of military innovation and adoption, as well as warfare writ large.

The Department of War is moving aggressively to field these capabilities, with the Secretary of War calling for AI adoption at "wartime speed." Congress is not on the same timeline. This paper makes the operational, risk-management case for why Congress needs to act, and why a statutory floor on human involvement in lethal engagement decisions is the minimum intervention required to preserve military effectiveness.

Key Findings

- **There are many advantages with autonomous lethal engagements, but those advantages contain great risk.** Autonomy delivers genuine value in data processing, command-level decision speed, and communications-denied environments. But those same benefits that make autonomy operationally necessary are the exact conditions under which the technical fragility of current AI models is most dangerous, automation bias in operators is most intense, and a human fail-safe is structurally absent.
- **Current AI/ML systems introduce forms of opacity, distributional fragility, and adversarial susceptibility that greatly differ from the failure modes of previous rule-based systems.** Machine-learning models are trained rather than programmed, and their decision rules are emergent. The performance outside the training distribution degrades unpredictably and without warning. A targeting algorithm trained on one operational environment will produce confident outputs in another that are also wrong, and neither operators nor current systems can be relied upon to identify these failures consistently in real time.
- **Adversaries will actively exploit this technology.** Data poisoning, adversarial inputs, electronic warfare, and behavioral manipulation turn an opponent's reliance on autonomy from an advantage into a targetable vulnerability.
- **"Human on the loop" oversight collapses under realistic operational conditions.** Automation bias, compressed decision windows, and the speed mismatch between machine operations and human cognition systematically transform nominal supervision into rubber-stamping.
- **The risk window is narrowing.** As autonomous capabilities proliferate and procurement commitments deepen, the structural conditions for humans to stay at pace with warfare become progressively harder to ensure retroactively.

---

[1] Ramkumar, Amrith, Keach Hagey, and Vera Bergengruen. "Pentagon Used Anthropic's Claude in Maduro Venezuela Raid." *Wall Street Journal*, February 15, 2026. https://www.wsj.com/politics/national-security/pentagon-used-anthropics-claude-in-maduro-venezuela-raid-583aff17.

[2] Copp, Tara, Elizabeth Dwoskin, and Ian Duncan. "Pentagon Leverages AI in Iran Strikes amid Feud with Anthropic." *Washington Post*, March 4, 2026. https://www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign/.

Policy Recommendations

- **Establish a statutory minimum decision window for autonomous lethal engagement.** Congress should codify a 5-second minimum human decision window before autonomous lethal engagement. This would include waiver authority for operational commanders and categorical exclusions for time-critical defensive systems. Without at least a few seconds to process, the human operator can collapse into a rubber-stamping function. A temporal minimum also ensures that semi-autonomous operation is the default and treats fully autonomous engagement as a mission-tailored exception.
- **Establish a procurement certification framework.** Congress should require AI systems to demonstrate privacy, robustness, oversight, verification, and explainability standards before contract award. These requirements will be documented through a standardized model card that reports assessments performed, representative training data, known failure modes, and appropriate use cases. The operational risks of brittleness, opacity, and adversarial vulnerability are upstream design problems, not operator problems. Once a system is fielded, the conditions for meaningful oversight are largely fixed. Certification forces those design choices to be surfaced and reviewable before contract award, giving the military and Congress a shared empirical basis for knowing where a system can be trusted and under what conditions it will most likely fail.

AI-Warfare is Coming to a Theater Near You

Autonomous and semi-autonomous weapon systems are not new. The United States and other states have fielded systems with autonomous functions for decades: the Phalanx Close-In Weapon System, the Navy's Aegis Combat System, the Army's Patriot air and missile defense system, and Harpy-class loitering munitions.[3] Each of these platforms can, under defined conditions, select and engage targets without a human pulling the trigger on each shot. The political and legal debate over lethal autonomous weapon systems (LAWS) has often proceeded as though autonomy in lethal systems were a future, emerging problem; operationally, it is a half-century-old reality.

The novelty is the technology that is currently underwriting autonomy in warfare. These older, deployed systems, such as the Aegis and Patriot, are algorithmic, using "straightforward, deterministic IF-THEN codes of the kind in use since the days of punch cards and vacuum tubes."[4] Their behavior is, in principle, fully specifiable: a given input produces a given output, and engineers can trace why. While these systems have produced tragic accidents, those failures occurred within an architecture whose logic was retrospectively legible. Modern machine-learning systems do not work this way. They are trained, not programmed, and their behavior under conditions outside their training distribution is not reliably predictable. The convergence of this fundamentally different class of software is being proposed to extend autonomous targeting across a much broader range of operational contexts, beyond constrained air-defense envelopes and into ground operations, urban environments, and contested multi-domain scenarios. This development warrants serious and immediate policy attention.

The technological discontinuity does not, on its own, settle the policy question. A serious case for AI-enabled autonomy in military systems exists. Autonomy, particularly AI-enabled autonomy, offers operational value across multiple use cases, from data processing, targeting, decision support, and more. AI processes information at scales and speeds no human team can match, which then compresses decision cycles in ways that materially affect tactical outcomes. These use cases provide advantages to a military if they enable conducting an Observe Orient Decide Act (OODA) loop faster than the adversary; AI helps speed up the observation and orientation, enabling quicker decisions and actions. Autonomy also allows platforms to continue functioning when communication with human controllers is degraded or severed. Each of these benefits are real and are already being exploited by the U.S. military. Autonomy delivers great value to the military, "including opportunities to reduce the number of warfighters in harm's way, increase the quality and speed of decisions in time-critical operations, and enable new missions that would otherwise be impossible."[5]

The first important use case is data processing. "Modern machine-learning AI can sort through much larger and more complicated datasets," [6] which is extremely valuable given the vast amount of data coming into a combatant command, an amount so large humans cannot possibly sort through it all.[7] This is more advanced than any prior generation of military software and is a capability whose value scales directly with the volume of sensor data flowing into a modern combatant command. This data processing scale and speed also

---

[3] Scharre, Paul. *Autonomous Weapons and Operational Risk*. Ethical Autonomy Project. Washington, DC: Center for a New American Security, February 2016. Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton, 2018.

[4] Freedberg, Sydney J., Jr. "Ethical Terminators, or How DoD Learned to Stop Worrying and Love AI: 2023 Year in Review." *Breaking Defense*, December 22, 2023. https://breakingdefense.com/2023/12/ethical-terminators-or-how-dod-learned-to-stop-worrying-and-love-ai-2023-year-in-review/.

[5] Defense Science Board. *Summer Study on Autonomy*. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, June 2016, p. 1.

[6] Freedberg, "Ethical Terminators."

[7] Huitt, Joseph L. "Leadership: Artificial Intelligence in Decision-Making." U.S. Army, July 3, 2025. https://www.army.mil/article/286847/leadership_artificial_intelligence_in_decision_making.

translates to clear tactical benefits. The greater scale enables an automated system to target a wider range of possible targets. Instead of being able to just track missiles and jets in the empty air, these machines can easily track additional targets such as hidden ground vehicles and people.

Project Maven/Maven Smart System, the DoD's flagship AI initiative launched in 2017, uses machine learning algorithms to process Intelligence, Surveillance, and Reconnaissance (ISR) drone footage. As a result, these systems can identify objects of interest significantly faster than human analysts, reducing the workload on intelligence personnel who were overwhelmed by the volume of full-motion video data they could not manually review.[8] Maven is the clearest available example of an autonomous function (automated target recognition) improving accuracy and reducing analyst error in a supporting role. It is also, notably, a non-lethal application with lethal implications. The algorithm surfaces targets, and humans decide what to do with them.

The debate around the increased target production has become extremely relevant in early 2026 with the war in Iran. Reporting has indicated the use of commercial AI system, primarily Anthropic's Claude embedded within Maven, in active operations against Iran; automated targeting pipelines have reportedly enabled what defense analysts identified the ability to strike a high volume of targets with precision-guided munitions at a tempo human-only targeting cycles could not sustain.[9] Recent reporting indicates AI-assisted targeting has accelerated target generation and operational tempo.[10] With this success and clearly established advantage, the question going forward is not whether to automate military functions but how to capture these gains without sacrificing the human judgment needed when autonomous systems fail or encounter conditions outside their training data. The same operations have also produced documented targeting errors. The conditions that make autonomous targeting operationally valuable (speed, volume, compressed decision cycles) are the same conditions under which human oversight degrades, and system errors propagate faster than they can be caught. These vital questions are already being asked[11] as targeting mistakes have been occurring in current operations against Iran and it remains unclear if human analysts have the ability and capacity to truly assess and intervene in AI targeting recommendations.

The second use case of operational value is decision quality for operational and tactical commanders. The speed of information processing and synthesis available to an AI-enabled command structure is faster than what a team of human operators can produce, "greatly accelerat[ing] the pace of information update" and compressing the interval between observation and commander awareness.[12] This helps to cut through the uncertainty and fog of war and reduces the cognitive burden on commanders managing complex, multi-domain operations. This is a central operational argument for autonomy. In a contested engagement, the side that can observe, orient, decide, and act faster than its adversary holds a structural advantage, and AI-enabled systems extend that advantage in ways human-only command architectures cannot match. The appeal is not speed for its own sake but the prospect of converting informational superiority into decision superiority before an adversary can exploit the gap.

---

[8] Defense Science Board, *Summer Study on Autonomy*, p. 49.
[9] Schneider, Jordan, Emmy Probasco, Michael Horowitz, Bryan Clark, and Henry Farrell. "Emergency Pod: Iran + Anthropic." *ChinaTalk* (podcast), March 2, 2026. https://www.chinatalk.media/p/emergency-pod-iran-anthropic.
[10] Horowitz, Michael C., and Paul Scharre. "Do Killer Robots Save Lives?" *Politico Magazine*, November 19, 2014. https://www.politico.com/magazine/story/2014/11/killer-robots-113010/.
Scharre, *Autonomous Weapons and Operational Risk*.
[11] Van Hollen, Chris, Tim Kaine, Elizabeth Warren, Brian Schatz, and Chuck Schumer, et al. Senators to Pete Hegseth, Secretary of Defense. March 11, 2026. Posted on the website of Senator Tammy Baldwin. https://www.baldwin.senate.gov/news/press-releases/baldwin-presses-trump-admin-for-answers-on-the-school-bombing-and-civilian-casualties-in-iran.
[12] Defense Science Board, *Summer Study on Autonomy*, p. 16.

The third and strongest operational case is the communications-denied environment. It makes sense to remove human oversight precisely when communications can be lost with human controllers. Communications-denied environments are "a primary driver as autonomous systems can continue operating when communications links are degraded or severed, a scenario that is 'especially relevant' in contested environments where adversaries will actively jam, spoof, or destroy communications infrastructure."[13] In such environments, autonomy is not a convenience or an efficiency gain; it is the difference between a platform that continues its mission and one that goes inert.[14] But these cases, where humans cede control and have no way to recall actions are when the risks of unintended engagements and escalation are highest, since the human fail-safe is absent.

This is the strongest operational case for removing humans from the immediate loop. It is also the case that most clearly exposes the problem. The advantages of autonomy that warrant operational necessity are precisely the conditions under which there is the greatest chance for fatal error. The operational argument for autonomy and the operational argument against fully autonomous lethal engagement is, in the end, built from the same set of facts.

In Case You Missed the Memo

The operational use cases are not theoretical, and the Department of War is not waiting for Congress to resolve the policy questions they raise. The Department has made a deliberate institutional judgment that the benefits of AI-enabled autonomy outweigh the risks of rapid adoption, and it is acting on that judgment at a pace that should inform how Congress thinks about the timeline for legislative action.

This priority is articulated across the Department's directives and memorandums. The 2023 Data, Analytics, and AI Adoption Strategy established AI integration as a priority.[15] The Secretary of War, Pete Hegseth, recently released a series of memos outlining an AI-first approach. Specifically, Secretary Hegseth called for the DoD to adopt AI capabilities with a "wartime speed".[16] He has also stated that the Pentagon needs to "accelerate like hell" in terms of AI adoption.[17] This strategy has been further supported by the aggressive tactics in AI procurement that the Department has used. In the recent Anthropic-Pentagon contract dispute, the Pentagon wanted to preserve full capabilities in the present, rather than agree to safety guardrails the AI company wanted to preserve.[18] These gestures and strategies are consistent with a sustained institutional push across two administrations to treat AI adoption as a capability race in which the cost of moving slowly is calculated as exceeding the cost of moving without fully resolved safety, reliability, and oversight frameworks.

Autonomy has great military benefits that AI's unique technology is further enabling to a degree never before seen. But the technology and all the operational benefits have clear and potentially dangerous tradeoffs. These

---

[13] Defense Science Board, *Summer Study on Autonomy*, p. 45, 49.
[14] Scharre, *Autonomous Weapons and Operational Risk, p. 45-46.*
[15] U.S. Department of Defense. *DoD Data, Analytics, and Artificial Intelligence Adoption Strategy.* Washington, DC: Department of Defense, November 2, 2023. https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF.
[16] U.S. Department of War. *Artificial Intelligence Strategy for the Department of War.* Washington, DC: Department of War, January 2026. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.
[17] Vincent, Brandi. "Pentagon Redefines Its Overarching Plan to Accelerate Data and AI Adoption," *DefenseScoop*, November 2, 2023, https://defensescoop.com/2023/11/02/pentagon-redefines-its-overarching-plan-to-accelerate-data-and-ai-adoption/.
[18] Perlo, Jared. "Anthropic Sues Trump Administration in AI Dispute with Pentagon." *NBC News*, March 9, 2026. https://www.nbcnews.com/tech/tech-news/anthropic-sues-trump-administration-ai-dispute-pentagon-rcna262444.

novel risks include unanticipated escalation and the erosion of human oversight, which will clearly threaten military effectiveness. AI-enabled AWS and other military systems are a categorically new type of weapon system that create novel and unique risks that demand distinct, regulatory actions.[19]

The implication for Congress is a matter of timing, not ideology. The greater the gap between adoption pace and governance pace, the harder it becomes to impose structural constraints retroactively. Programs in procurement acquire constituencies; fielded systems acquire operational dependencies; doctrine adapts to available capabilities rather than the reverse. The window during which Congress can establish baseline human-oversight requirements as a condition of fielding rather than as a retrofit to systems already in the force is not indefinite, and on the trajectory the Department has set, it is narrowing.

The Technology is Not There Yet

Current AI relies on specific technologies such as machine learning and neural networks, and the nature of these technologies produce uncertainty as they are not deterministic algorithms. Modern machine-learning systems differ along three dimensions that bear directly on their suitability for lethal applications in the military. Decision rules are learned rather than specified, performance outside of training distribution degrades in ways designers cannot reliably predict, and failures arrive without the warning signs that operators of earlier generations of military software were trained to recognize. Neural networks do not perform rule-based calculations; they learn by exposure to large data sets and output predictions. Accordingly, "the internal structure of the network that generates output can be opaque to the designers known as the black box."[20] This opacity is a structural consequence of how the technology works, and it has three consequences that matter for weapons applications: the systems are uninterpretable, they are unreliable as errors cannot be predicted from inputs, and they are brittle in ways that older algorithmic software is not.

One consequence for military operations is distributional shift.[21] Machine-learning systems perform well on inputs that resemble their training data and degrade, often rapidly, on inputs that do not. When operational conditions differ from training data, ML model accuracy degrades rapidly and unpredictably. For example, a model trained on overhead imagery from Houston that would not perform well on imagery from Russia or Germany, where landscapes differ significantly.[22] There is great risk in deploying targeting algorithms trained on one theater's conditions into a different operational environment.

A targeting algorithm trained on data from one operational environment will, when deployed in a different one, produce confident outputs that are also wrong, and it will produce them in ways neither the operator nor the system itself can flag in real time. This is not a hypothetical concern. Maven's computer vision component initially only had representational training data from the Middle East, specifically the U.S.'s time in Afghanistan, and when the technology was being used to monitor refugee movement in Ukraine when Russia first invaded, it became clear that Maven was not going to be useful; as a result, satellite images were taken of

---

[19] Simmons-Edler, Riley, Jean Dong, Paul Lushenko, Kanaka Rajan, and Ryan P. Badman. "Military AI Needs Technically-Informed Regulation to Safeguard AI Research and Its Applications." arXiv preprint arXiv:2505.18371v2, November 11, 2025. https://arxiv.org/abs/2505.18371.

[20] Scharre, *Autonomous Weapons and Operational Risk, p. 15.*

[21] Simmons-Edler, et al. "Military AI Needs Technically-Informed Regulation to Safeguard AI Research and its Applications."

[22] Hoffman, Wyatt, and Heeu Millie Kim. *Reducing the Risks of Artificial Intelligence for Military Decision Advantage.* Washington, DC: Center for Security and Emerging Technology, March 2023. https://cset.georgetown.edu/wp-content/uploads/CSET-Reducing-the-Risks-of-Artificial-Intelligence-for-Military-Decision-Advantage.pdf.

Ukraine's operation environment to add this needed representation to the model's training data.[23] This is not a failure of the algorithm in the way a software bug is a failure; it is the algorithm functioning exactly as designed, on inputs it was not designed for.

This fragility also contributes to a fundamental problem of "brittle competence," where autonomous systems that can perform reliably under assigned, normal operating conditions, but "if pushed beyond the bounds of their programming, they may fail and fail badly."[24] The problem intensifies with system complexity. As the operational envelope expands and the number of variables in the input space grows, the region within which the system is reliable becomes harder for operators and designers to characterize in advance. The boundary between reliable and catastrophic performance can be crossed without anyone knowing it has been crossed until the consequences manifest. Failures arrive without warning.

There are additional risks when using this technology in combat. Adversaries will not passively allow autonomous systems to function as designed. Electronic warfare, spoofing, adversarial inputs, and behavioral manipulation represent force multipliers against autonomous targeting, turning an opponent's reliance on autonomy into a vulnerability from an advantage. Adversarial risks are a qualitatively different problem from this. What is clear is that this technology is being used and that "when the behavior of the autonomous system can be predicted, it is susceptible to behavioral hacking by adversaries."[25] Anywhere a machine-learning system's outputs are a reproducible function of its inputs, a sufficiently sophisticated adversary can engineer inputs that produce outputs the operator does not want. Military AI operates in environments where the adversarial actor is a peer intelligence service with the resources to study the system, discover its failure modes, and trigger them deliberately.

There are many ways an adversary can tamper with AI models, but there are two specific attack types that are especially concerning. The first is "data poisoning." Adversaries subtly manipulate the datasets used to train targeting algorithms so that systematic misclassifications are baked in before deployment.[26] A model trained on poisoned data will consistently misidentify certain object classes (e.g., civilian vehicles as military targets, or the reverse). It is unclear how sufficient and robust current provenance and verification practices are, especially for the high-risk, critical military applications.[27] Neither the operator nor the system will have any means of detecting the corruption in real time because the system is functioning exactly as its corrupted training taught it to function.[28]

The second is adversarial input manipulation at the point of use, where small physical modifications to objects in the environment produce outsized changes in how the system classifies them. AI can "fail due to adversarial attacks intentionally designed to trick or fool algorithms into making a mistake."[29] Researchers have demonstrated, for example, that pieces of adhesive tape placed on a road sign in specific patterns can cause an image-recognition system to confidently misclassify a stop sign as a speed limit sign.[30] Other demonstrated attacks include image classification systems deceived by pixel-level changes into misidentifying

---

[23] Allen, Gregory C. "Inside Project Maven and AI-Powered Warfare with Katrina Manson." *AI Policy Podcast* (podcast), Center for Strategic and International Studies, March 26, 2026.. https://www.csis.org/podcasts/ai-policy-podcast/inside-project-maven-and-ai-powered-warfare-katrina-manson.

[24] Scharre, *Autonomous Weapons and Operational Risk, p. 6-7, 15-17*.

[25] Scharre, *Autonomous Weapons and Operational Risk, p. 36-37*.

[26] Hoffman and Kim. "Reducing the Risks of Artificial Intelligence for Military Decision Advantage."

[27] Banerjee, Dave. "Securing AI Infrastructure to Prevent Backdoors and Sabotage." *The Substrate* (Substack blog), March 26, 2026., https://www.the-substrate.net/p/securing-ai-infrastructure-to-prevent.

[28] Pedersen, Joel. "Architecting Trust: A Modular Framework for the Operational Deployment of Autonomous Systems." *Technology & National Security Review* 1, 2026.

[29] Joint Air Power Competence Centre. *Adversarial Machine Learning*. Kalkar, Germany: Joint Air Power Competence Centre, July 2022. https://www.japcc.org/essays/adversarial-machine-learning/.

[30] JAPCC, *Adversarial Machine Learning*.

objects, game-playing AI that collapse when rules are slightly altered, and autonomous vehicles induced to swerve into oncoming traffic by small pieces of tape on road signs. As a result, "keeping humans in or on the loop is essential" because humans can recognize adversarial attacks that algorithms cannot.[31]

The inherent limitations of current AI technology and the application in an adversarial, combat environment is a compounding problem. Distributional shift means that any deployment outside the training environment introduces uncertainty that the system cannot flag. Brittle competence means that the transition from working to failing will not come with warning. Adversarial manipulation means that a capable opponent will actively work to push the system across that boundary in ways neither the operator nor the designer has anticipated. Any one of these properties, taken alone, would be a serious argument for retaining human oversight of lethal engagement decisions. Taken together, they establish that the technology in its current form cannot be relied upon to make those decisions without a human in a position to recognize and interrupt the failures it will inevitably produce.

Where Do Humans Fit

Human oversight can only do so much to mitigate the inherent and adversarial risks mentioned above. The architecture of an often-proposed AI-enabled weapon assumes exactly this: a human "on the loop," monitoring autonomous outputs and intervening when something goes wrong. The assumption does not survive contact with operational reality.

Human oversight of autonomous weapons systems degrades predictably under operational conditions. Automation bias distorts operator judgment, compressed decision timelines eliminate deliberation, and the historical record of autonomous and semi-autonomous systems already in the force shows that "human on the loop" has repeatedly failed to prevent the exact class of catastrophic errors it was designed to prevent. The problem is not the absence of a human in the kill chain. It is the structural conditions of contemporary combat that render the human's presence functionally meaningless.

The degree of human involvement in weapon systems exists on a spectrum with different framings, from fully manual to fully autonomous. Leading experts Scharre and Horowitz described the spectrum with three control types: semi-autonomous ("human in the loop," where the system stops and waits for human approval), supervised autonomous ("human on the loop," where the system acts under human supervision with ability to intervene), and fully autonomous ("human out of the loop," where the human cannot observe and correct in sufficient time).[32] This is distinctly different from automated, which means a deterministic algorithm is in use. The critical variable is the time between system failure and human corrective action, which changes dramatically depending on control type.[33] The current policy informing the U.S. defense institutions, DoDD 3000.09, defines the categories based on human involvement in targeting.[34] The directive defines autonomous weapon systems as systems that "once activated, can select and engage targets without further intervention by a human operator," while semi-autonomous systems as systems requiring a human to "select and engage specific targets."[35]

---

[31] JAPCC, *Adversarial Machine Learning*.

[32] Scharre, Paul, and Michael C. Horowitz. *An Introduction to Autonomy in Weapon Systems.* CNAS Working Paper. Washington, DC: Center for a New American Security, February 2015. https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.
Scharre, *Autonomous Weapons and Operational Risk,* p. 8-10.

[33] Scharre and Horowitz, *Introduction to Autonomy*. Scharre, *Autonomous Weapons and Operational Risk,* p. 8-10.

[34] U.S. Department of Defense. *Autonomy in Weapon Systems.* DoD Directive 3000.09. Washington, DC: Department of Defense, January 25, 2023. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

[35] DoD Directive 3000.09, Glossary.
Congressional Research Service. *Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems.* IF11150. Washington, DC: Congressional Research Service, https://www.congress.gov/crs-product/IF11150.

FIGURE 1 · CONCEPTUAL FRAMEWORK

The Spectrum of Human Involvement in Weapon Systems

Human control over weapon systems exists on a continuum. Leading scholars Paul Scharre and Michael Horowitz organize this continuum into three categories of control, each defined by where the human sits relative to the decision loop.

CATEGORY 01

Semi-Autonomous

Human in the loop

The system stops and waits for explicit human approval before engaging. The operator authorizes each action.

CATEGORY 02

Supervised Autonomous

Human on the loop

The system acts on its own under active human supervision, with the operator retaining the ability to intervene.

CATEGORY 03

Fully Autonomous

Human out of the loop

The system selects and engages targets without the operator being able to observe and correct in sufficient time.

POLICY FRAMING **DoDD 3000.09 categorizes by human involvement in targeting**

SEMI-AUTONOMOUS WEAPON SYSTEM

A system that requires a human to **select and engage specific targets.**

AUTONOMOUS WEAPON SYSTEM

A system that, once activated, **can select and engage targets without further intervention** by a human operator.

The central problem with "human on the loop" as an oversight model is that it performs well on paper and poorly in practice, and the reasons for this are structural rather than incidental. Under realistic operational conditions, the human on the loop model collapses into rubber-stamping. One reason is automation bias. In high-tempo operations, operators nominally in control exhibit automation bias due to an "unwarranted" and uncritical trust in automation.[36] The consequence is not merely suboptimal decision-making but a functional transfer of targeting authority from the human to the machine; operators approve rather than decide.[37]

Air-defense history suggests a recurring pattern: under high tempo and compressed timelines, nominal supervision can degrade into ratification. Modern AI systems will intensify that dynamic. The second Patriot fratricide in Operation Iraqi Freedom is an illustration of human ratification instead of human-machine teaming. The accident occurred *after* the Army had already implemented a corrective measure designed to restore human control following the Tornado incident; dropping launchers to standby so the system would require a deliberate human step before engaging.[38] Yet when the tactical director ordered units to bring launchers to ready status, that command, intended only as preparation, automatically triggered an engagement. The tactical director "either did not know that, or he did not remember in the heat of impending action that returning launchers to ready status would result in an automatic engagement."[39] The human was nominally in control; but operationally, the machine decided. Decades of deploying air defense systems with automated and autonomous features have incrementally normalized a diminished human role in specific use-of-force decisions, creating an emerging norm in which operators function as ratifiers of machine outputs rather than independent decision-makers. Existing systems have rendered human control "meaning-less" in practice even where it exists in doctrine.[40]

---

[36] Scharre, *Autonomous Weapons and Operational Risk*.

[37] Scharre, *Autonomous Weapons and Operational Risk*.

[38] Hawley, John K. *Patriot Wars: Automation and the Patriot Air and Missile Defense System.* Washington, DC: Center for a New American Security, January 2017. https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Report-EthicalAutonomy5-PatriotWars-FINAL.pdf. P. 7-8.

[39] Hawley, *Patriot Wars*, p. 7-8.

[40] Bode, Ingvild, and Tom Watts. *Meaning-less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS*. Odense: Center for War Studies, University of Southern Denmark, February 2021. https://pure.royalholloway.ac.uk/en/publications/meaning-less-human-control-lessons-from-air-defence-systems-on-me/.

Compression of decision timelines contributed to the automation bias. Research on time-critical automated decision support systems establishes that as engagement timelines compress, automation bias intensifies.[41] Operators under time pressure are systematically less likely to scrutinize system recommendations and more likely to accept them without independent verification. Military command and control is a high-risk domain precisely because of the combination of time pressure, high stakes, and system complexity that characterizes combat environments. In such conditions, higher levels of automation are "not advisable because of the risks and the complexity of both the system and the inability of the automated decision aid to be perfectly reliable."[42] To mitigate this effect, operators need time and training to exercise judgment and be a part of a culture that emphasizes their responsibility to ensure their role is effective and meaningful.

The Patriot engagement decision window illustrates the concrete military consequence: in the Tornado fratricide, the operator had less than ten seconds from detection to kinetic contact.[43] This timeline was not an aberration for this type of operation, but it was precisely this compression that made the distinction between automatic and manual modes operationally meaningless. An operator who cannot cognitively process, contextualize, and override a system recommendation within the available window is not exercising control. They are witnessing an automated decision.

As a result, more participatory human-machine teaming with a basis of humans in the loop is valuable operationally. But this type of human role requires certain parameters to be effective and substantive. The mismatch between machine operating speeds and human cognitive bandwidth structurally degrades decision quality independent of operator training or intent. Speed is not a neutral technical parameter but the primary mechanism by which autonomy functionally displaces human judgment regardless of formal command architecture.[44] Fully autonomous is more risky than semi-autonomous as fully AWS "have a qualitatively different degree of risk than equivalent semi-autonomous weapons that would retain a human in the loop," and the consequences are far greater with autonomy.[45] These failures could range from fratricide, civilian casualties, or inadvertent & accidental escalation in a crisis.

Policy Recommendations

The risks of fully autonomous lethal engagement are not addressable through doctrine alone. Congress should establish a statutory floor for human oversight through mandatory decision-time thresholds, paired with procurement certification requirements, to stress-test AI-enabled operations. The Human in the Loop Delay Minima proposal represents the minimum structurally necessary intervention: not a guarantee of meaningful oversight, but the minimal precondition without which meaningful oversight is impossible. It constrains fully autonomous functions to semi-autonomous functions.

Absent humans in the loop involvement and control, or carefully justified exceptions, the operational advantages of AI can easily be outweighed by exploitable failure modes in lethal decision contexts. The operational evidence presented in this paper establishes that the risks of fully autonomous lethal engagement are not addressable through existing DoD policy.

---

[41] Cummings, M. L. "Automation Bias in Intelligent Time Critical Decision Support Systems." Paper presented at the AIAA 1st Intelligent Systems Technical Conference, Chicago, IL, September 20–22, 2004. https://doi.org/10.2514/6.2004-6313.
Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association* 19, no. 1 (January 2012): 121–27. https://doi.org/10.1136/amiajnl-2011-000089.

[42] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems."

[43] Hawley, *Patriot Wars*.

[44] Scharre, *Autonomous Weapons and Operational Risk,* p. 5.

[45] Scharre, *Autonomous Weapons and Operational Risk,* p. 5.

The case for retaining human oversight is operationally grounded. Scharre states:

> "In fact, for many applications, it may still be possible to keep a human in the loop in a semi-autonomous mode of operation without sacrificing much in the way of a time delay. Humans would not need to physically maneuver the weapon — after all, homing missiles and torpedoes are 'fire and forget' today — but only remain in charge of authorizing targets for engagement."[46]

There are conditions under which human oversight is operationally meaningful rather than nominal: effective human decision-making requires that operators and commanders have a functional understanding of the targeting system, situational understanding of the engagement context, and the capacity to scrutinize machine outputs rather than defer to them[47]. These conditions require deliberate structural support through interface design, training, and minimal time to build situational awareness.

Ensuring Enough Time for Humans to Be "In the Loop"

Congress should establish a statutory floor for ensuring effective autonomy in the military through two reinforcing mechanisms: mandatory human decision-time thresholds for autonomous lethal engagement and a procurement certification framework requiring AI systems to demonstrate reliability and interpretability before fielding.

The Human in the Loop Delay Minima proposal codifies a 5-second minimum decision window before autonomous lethal engagement. This provision will include a mission necessary waiver authority for operational commanders subject to congressional notification, as well as categorical exclusions for time-critical defensive systems where any latency is incompatible operationally. This structure mirrors existing congressional notification frameworks for nuclear posture, cyber operations, and covert action, treating autonomous lethal engagement as a category of action warranting equivalent institutional oversight.[48]

The 5-second threshold is not arbitrary; it is a deliberately conservative minimum intended to preserve some possibility of intervention in fast-paced engagements. Cognitive science literature on situational awareness in time-critical decision environments establishes that operators require a minimum window to perceive a system recommendation, assess the target and context, and intervene if the machine output is wrong.[49] Below roughly 3 seconds, people cannot build even minimally effective situational awareness; meaningful gains in comprehension continue up to approximately 7 seconds, after which additional time yields diminishing returns. At engagement timelines of 10 seconds or less, the conditions under which the Patriot fratricides in OIF occurred,[50] the distinction between automatic and manual modes become meaningless as the operator cannot cognitively process and override a recommendation within the available window. The 5-second floor

---

[46] Scharre, *Autonomous Weapons and Operational Risk,* p. 46.

[47] Bode and Watts, *Meaning-less Human Control.*

[48] Fink, Anya L. *Congressional Commission on the U.S. Strategic Posture.* CRS In Focus IF12621. Washington, DC: Congressional Research Service, December 10, 2025. https://www.congress.gov/crs-product/IF12621.
U.S. Code. Title 10, § 394. https://www.law.cornell.edu/uscode/text/10/394.
DeVine, Michael E. *Covert Action and Clandestine Activities of the Intelligence Community: Selected Definitions.* CRS Report R45175. Washington, DC: Congressional Research Service, November 29, 2022. https://www.congress.gov/crs-product/R45175.

[49] Lu, Zhenji, Xander Coster, and Joost de Winter. "How Much Time Do Drivers Need to Obtain Situation Awareness? A Laboratory-Based Study of Automated Driving." *Applied Ergonomics* 60 (April 2017): 293–304. https://doi.org/10.1016/j.apergo.2016.12.003.
Bryant, David J., and David G. Smith. *Comparison of Identify-Friend-Foe and Blue-Force Tracking Decision Support for Combat Identification.* Technical Report. Toronto: Defence Research and Development Canada, February 2011. https://www.researchgate.net/publication/235080213.

[50] Hawley, *Patriot Wars.*

sits at the lower bound of the range within which human judgment is functionally possible, not optimal.

This delay minima by itself is not a guarantee of meaningful oversight, but it is the first of structural preconditions to make human judgement actionable. Time is a necessary but insufficient condition: operators also require training, interface design that surfaces the basis for machine recommendations, and a command culture that reinforces their responsibility to scrutinize rather than defer to system outputs.[51] What the delay minima does accomplish is to establish the first structural precondition without which the others cannot function. No amount of training or interface design can compensate for a decision window too compressed for human cognition to operate within it.

The threshold also performs a second function as it establishes semi-autonomous operation as the statutory baseline and treats fully autonomous lethal engagement as a mission-tailored exception requiring affirmative justification through the waiver or categorical exclusion process. A statutory floor forces the burden of justification onto autonomy rather than onto human control, ensuring that removing humans from the decision is a deliberate choice out of mission necessity.

Establish Reliable Procurement Certifications

Additionally, a procurement certification process focused on AI reliability and appropriate military use cases addresses risks that doctrine and operator training cannot reach. Once a system is fielded, the structural conditions for meaningful oversight are largely fixed by choices made prior to fielding. These include what data the model was trained on, how its failure modes were characterized, and what operational envelope it was validated against. Certification moves the point of intervention earlier, requiring AI systems to demonstrate privacy, robustness, oversight, verification, and explainability standards before contract award rather than discovering their limitations in combat.

These standards would be recorded in a standardized, defense-tailored model card accompanying each certified system. Unlike the voluntary model cards used in commercial AI, a defense model card would be a required deliverable specifying the representative training data and its known gaps, the operational envelopes within which the model has been validated, the conditions under which performance is expected to degrade, and documented failure modes from testing and evaluation. Critically, it would specify not only appropriate use cases but inappropriate ones: the operational contexts, environments, or mission profiles for which the system is not certified and should not be fielded. The model card makes that boundary explicit rather than leaving it to be discovered operationally.

The certification is not a vehicle for tech companies to dictate policy or constrain operational flexibility. It is a mechanism for ensuring that program managers, operators, and Congress share a common basis for understanding what a system can and cannot do. Without this baseline, fielding decisions and rules of engagement are made without the information needed to exercise meaningful judgment.

Open Questions that Demand Congressional Action

The risks documented in this paper are not hypothetical. The question before Congress is not about permitting autonomous weapons, they already exist and have been for decades. This is about the new phase of technology and warfare meeting and addressing the real operational harm AI/ML can cause. Congress needs to decide whether to establish the minimum institutional conditions under which human judgment remains a real rather than nominal feature of lethal force decisions.

---

[51] Bode and Watts, *Meaning-less Human Control*.

Bibliography

Allen, Gregory C. "Inside Project Maven and AI-Powered Warfare with Katrina Manson." *AI Policy Podcast* (podcast). Center for Strategic and International Studies, March 26, 2026. https://www.csis.org/podcasts/ai-policy-podcast/inside-project-maven-and-ai-powered-warfare-katrina-manson.

Banerjee, Dave. "Securing AI Infrastructure to Prevent Backdoors and Sabotage." *The Substrate* (Substack blog), March 26, 2026. https://www.the-substrate.net/p/securing-ai-infrastructure-to-prevent.

Bode, Ingvild, and Tom Watts. *Meaning-less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS*. Odense: Center for War Studies, University of Southern Denmark, February 2021. https://pure.royalholloway.ac.uk/en/publications/meaning-less-human-control-lessons-from-air-defence-systems-on-me/.

Boulanin, Vincent, Tytti Erästö, Petr Topychkanov, Lora Saalman, and Moa Peldán Carlsson. *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control.* Stockholm: SIPRI, June 2020.

Boulanin, Vincent, and Maaike Verbruggen. *Mapping the Development of Autonomy in Weapon Systems.* Stockholm: SIPRI, November 2017.

Brennan-Marquez, Kiel, Karen Levy, and Daniel Susser. "Strange Loops: Apparent versus Actual Human Involvement in Automated Decision-Making." *Berkeley Technology Law Journal* 34 (2019): 745–71.

Bryant, David J., and David G. Smith. *Comparison of Identify-Friend-Foe and Blue-Force Tracking Decision Support for Combat Identification.* Technical Report. Toronto: Defence Research and Development Canada, February 2011. https://www.researchgate.net/publication/235080213.

Congressional Research Service. *Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems.* CRS In Focus IF11150. Washington, DC: Congressional Research Service. https://www.congress.gov/crs-product/IF11150.

Copp, Tara, Elizabeth Dwoskin, and Ian Duncan. "Pentagon Leverages AI in Iran Strikes amid Feud with Anthropic." *Washington Post*, March 4, 2026. https://www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign/.

Crootof, Rebecca, Margot E. Kaminski, and W. Nicholson Price II. "Humans in the Loop." *Vanderbilt Law Review* 76 (2023): 429–510.

Cummings, M. L. "Automation Bias in Intelligent Time Critical Decision Support Systems." Paper presented at the AIAA 1st Intelligent Systems Technical Conference, Chicago, IL, September 20–22, 2004. https://doi.org/10.2514/6.2004-6313.

Defense Science Board. *Summer Study on Autonomy.* Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, June 2016.

DeVine, Michael E. *Covert Action and Clandestine Activities of the Intelligence Community: Selected Definitions.* CRS Report R45175. Washington, DC: Congressional Research Service, November 29, 2022. https://www.congress.gov/crs-product/R45175.

Fink, Anya L. *Congressional Commission on the U.S. Strategic Posture.* CRS In Focus IF12621. Washington, DC: Congressional Research Service, December 10, 2025. https://www.congress.gov/crs-product/IF12621.

Freedberg, Sydney J., Jr. "Ethical Terminators, or How DoD Learned to Stop Worrying and Love AI: 2023 Year in Review." *Breaking Defense*, December 22, 2023. https://breakingdefense.com/2023/12/ethical-terminators-or-how-dod-learned-to-stop-worrying-and-love-ai-2023-year-in-review/.

Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association* 19, no. 1 (January 2012): 121–27. https://doi.org/10.1136/amiajnl-2011-000089.

Hawley, John K. *Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons.* ARL-SR-0158. Adelphi, MD: U.S. Army Research Laboratory, September 2007.

Hawley, John K. *Patriot Wars: Automation and the Patriot Air and Missile Defense System.* Washington, DC: Center for a New American Security, January 2017. https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Report-EthicalAutonomy5-PatriotWars-FINAL.pdf.

Heginbotham, Eric, et al. *The U.S.-China Military Scorecard: Forces, Geography, and the Evolving Balance of Power, 2000–2017.* Santa Monica, CA: RAND Corporation, 2015.

Hoffman, Wyatt, and Heeu Millie Kim. *Reducing the Risks of Artificial Intelligence for Military Decision Advantage.* Washington, DC: Center for Security and Emerging Technology, March 2023. https://cset.georgetown.edu/wp-content/uploads/CSET-Reducing-the-Risks-of-Artificial-Intelligence-for-Military-Decision-Advantage.pdf.

Horowitz, Michael C. "Why Words Matter: The Real World Consequences of Defining Autonomous Weapon Systems." *Temple International and Comparative Law Journal* 30, no. 1 (2016): 85–98.

Horowitz, Michael C., and Paul Scharre. "Do Killer Robots Save Lives?" *Politico Magazine*, November 19, 2014. https://www.politico.com/magazine/story/2014/11/killer-robots-113010/.

Huitt, Joseph L. "Leadership: Artificial Intelligence in Decision-Making." U.S. Army, July 3, 2025. https://www.army.mil/article/286847/leadership_artificial_intelligence_in_decision_making.

Joint Air Power Competence Centre. *Adversarial Machine Learning.* Kalkar, Germany: Joint Air Power Competence Centre, July 2022. https://www.japcc.org/essays/adversarial-machine-learning/.

Lu, Zhenji, Xander Coster, and Joost de Winter. "How Much Time Do Drivers Need to Obtain Situation Awareness? A Laboratory-Based Study of Automated Driving." *Applied Ergonomics* 60 (April 2017): 293–304. https://doi.org/10.1016/j.apergo.2016.12.003.

Morgan, Forrest E., et al. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. Santa Monica, CA: RAND Corporation, 2020.

Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52, no. 3 (June 2010): 381–410.

Pedersen, Joel. "Architecting Trust: A Modular Framework for the Operational Deployment of Autonomous Systems." *Technology & National Security Review* 1 (2026).

Perlo, Jared. "Anthropic Sues Trump Administration in AI Dispute with Pentagon." *NBC News*, March 9, 2026. https://www.nbcnews.com/tech/tech-news/anthropic-sues-trump-administration-ai-dispute-pentagon-rcna262444.

Ramkumar, Amrith, Keach Hagey, and Vera Bergengruen. "Pentagon Used Anthropic's Claude in Maduro Venezuela Raid." *Wall Street Journal*, February 15, 2026. https://www.wsj.com/politics/national-security/pentagon-used-anthropics-claude-in-maduro-venezuela-raid-583aff17.

Saxon, Dan. "A Human Touch: Autonomous Weapons, DoD Directive 3000.09 and the Interpretation of 'Appropriate Levels of Human Judgment' over the Use of Force." In *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, and Claus Kreß, 185–208. Cambridge: Cambridge University Press, 2016. https://doi.org/10.1017/CBO9781316597873.009.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton, 2018.

Scharre, Paul. *Autonomous Weapons and Operational Risk*. Ethical Autonomy Project. Washington, DC: Center for a New American Security, February 2016.

Scharre, Paul, and Michael C. Horowitz. *An Introduction to Autonomy in Weapon Systems*. CNAS Working Paper. Washington, DC: Center for a New American Security, February 2015. https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.

Schneider, Jordan, Emmy Probasco, Michael Horowitz, Bryan Clark, and Henry Farrell. "Emergency Pod: Iran + Anthropic." *ChinaTalk* (podcast), March 2, 2026. https://www.chinatalk.media/p/emergency-pod-iran-anthropic.

Simmons-Edler, Riley, Jean Dong, Paul Lushenko, Kanaka Rajan, and Ryan P. Badman. "Military AI Needs Technically-Informed Regulation to Safeguard AI Research and Its Applications." arXiv preprint arXiv:2505.18371v2, November 11, 2025. https://arxiv.org/abs/2505.18371.

U.S. Code. Title 10, § 394. https://www.law.cornell.edu/uscode/text/10/394.

U.S. Department of Defense. *Autonomy in Weapon Systems*. DoD Directive 3000.09. Washington, DC: Department of Defense, January 25, 2023. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

U.S. Department of Defense. *DoD Data, Analytics, and Artificial Intelligence Adoption Strategy*. Washington, DC: Department of Defense, November 2, 2023. https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF.

U.S. Department of War. *Artificial Intelligence Strategy for the Department of War*. Washington, DC: Department of War, January 2026. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.

Van Hollen, Chris, Tim Kaine, Elizabeth Warren, Brian Schatz, Chuck Schumer, et al. Senators to Pete Hegseth, Secretary of Defense. March 11, 2026. Posted on the website of Senator Tammy Baldwin. https://www.baldwin.senate.gov/news/press-releases/baldwin-presses-trump-admin-for-answers-on-the-school-bombing-and-civilian-casualties-in-iran.

Vincent, Brandi. "Pentagon Redefines Its Overarching Plan to Accelerate Data and AI Adoption." *DefenseScoop*, November 2, 2023. https://defensescoop.com/2023/11/02/pentagon-redefines-its-overarching-plan-to-accelerate-data-and-ai-adoption/.